BD2K Investments in Training

Executive Summary

Training is a major limiting factor to extracting knowledge from data, and therefore it is a significant part of the Big Data to Knowledge (BD2K) Initiative. The primary goals of BD2K in training are to increase the number of biomedical data scientists and to improve the data science skills of all biomedical scientists.

Because training needs vary greatly based on an individual's prior background and intended use for data science, the BD2K investments in training are also varied. For those individuals who are primarily biomedical scientists and do not intend to become specialists in data science, BD2K supports courses and educational resources that are meant to enable participants to become conversant in data science and attain skills to utilize data science methods. To support those individuals who wish to become specialists in biomedical data science, BD2K includes training programs for predoctoral students, research rotations for early career scientists, and career development awards for postdocs and more senior researchers. To foster the development of new teams consisting of biomedical scientists and data scientists, BD2K is supporting the QuBBD (Quantitative Approaches to Biomedical Big Data) Program along with the National Science Foundation.

Among the first BD2K awards issued (in September 2014) were training awards, and they are starting to bear fruit. For example, the first BD2K Open Educational Resources, in the form of Massive Open Online Courses, have been released and already boast thousands of graduates; three summer courses were offered and filled beyond capacity; and individuals supported by career development awards have transitioned to more secure positions. Although the existing awards made in FY14-16 are promising, additional investments are needed to keep up with the fast-changing area of data science.

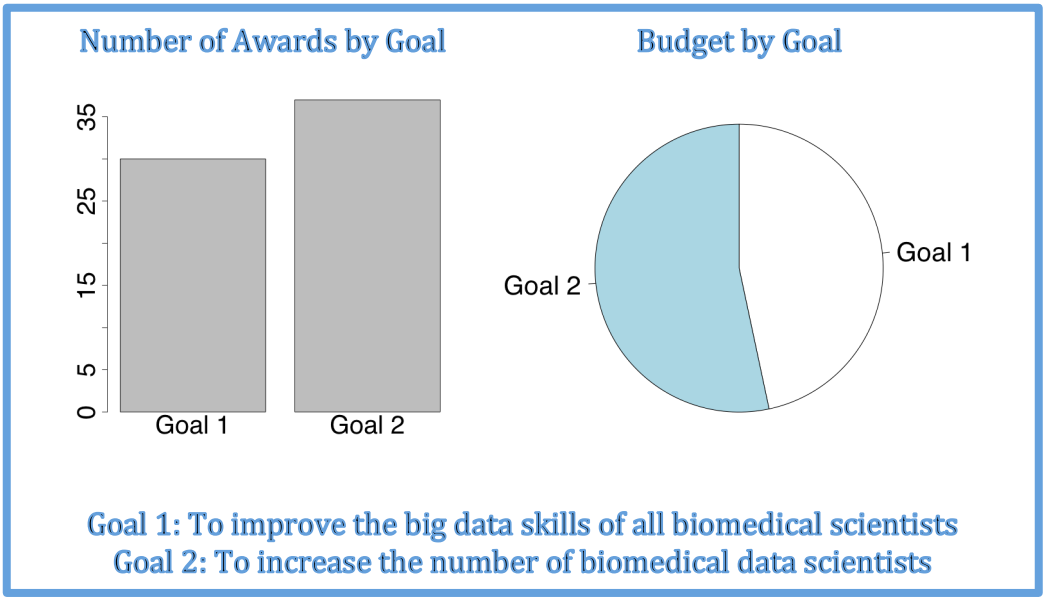
In order for BD2K-supported resources to have maximal impact, they need to be findable, accessible, interoperable, and reusable (FAIR). To help biomedical scientists find and access the most appropriate data science educational resources, the BD2K Training Coordination Center (TCC) is developing an Educational Resource Discovery Index, working with international partners. Through the TCC and the other training awards, BD2K aims to improve the ability of the entire biomedical science community, whether specialists in biomedical science or data science, to utilize the growing volume and complexity of data.

Overall Goals

Focusing on training was one of the main recommendations from the June 2012 Data and Informatics Working Group Report. It is also one of the major thrust areas of the BD2K program and the ADDS office. Training currently accounts for 15% of the BD2K budget and is expected to ramp to about 20%. The term "Training" is meant to encompass training, education, and workforce development that provides learners, no matter what career level, either foundational knowledge or skills for immediate use.

There are two main goals for training, to improve big data skills in all biomedical scientists and to increase the number of people who specialize in biomedical data science. These two goals include sub-goals as well:

1) To improve big data skills of all biomedical scientists
    a. Support training opportunities, both in-person and online
    b. Ensure training opportunities and resources are more readily discovered and accessed
    c. Enhance diversity in the biomedical and biomedical data science workforces
2) To increase the number of biomedical data scientists
    a. Establish biomedical data science as a career path
    b. Foster collaborations between biomedical scientists and data scientists



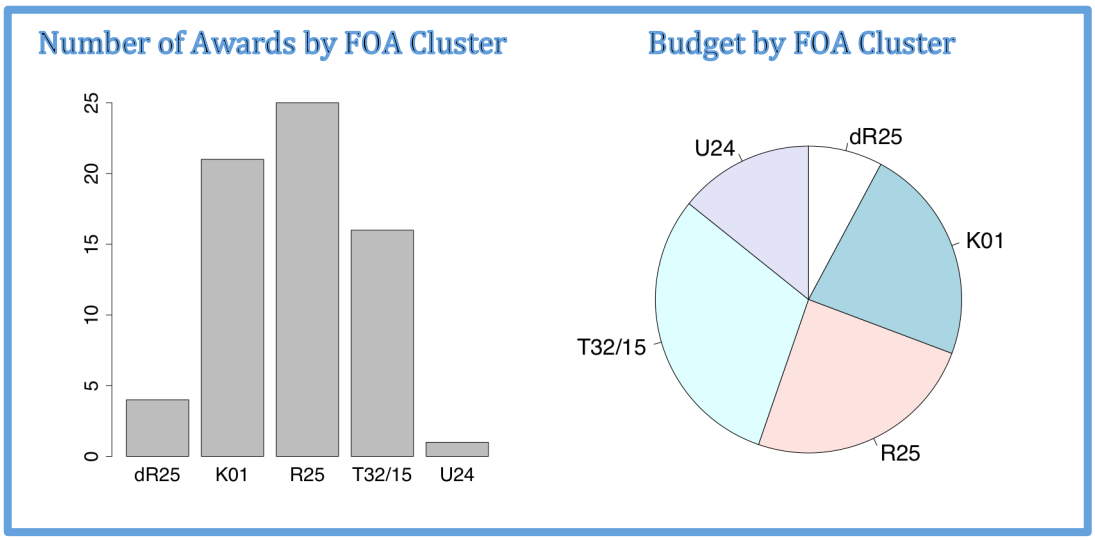
To accomplish these goals, the training portfolio is diverse. Although ten FOAs were issued in FY14 and FY15, they cluster into five groups:

- R25 awards for courses and resources about data management and data science (4 FOAs were issued to allow for different budgetary categories and structures)

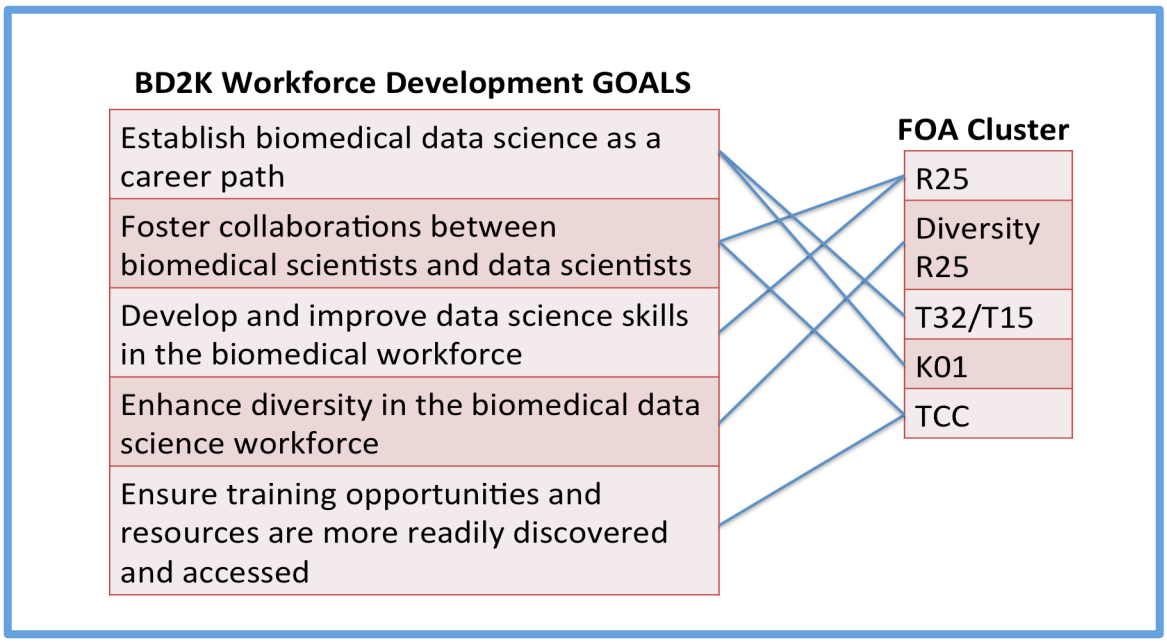
- R25 awards to enhance diversity (denoted dR25 to distinguish it from the other R25s)
- T32/T15 training programs (3 FOAS to support both new programs and supplements to existing T32s and T15s)
- K01 Career Development award
- U24 Training Coordination Center (TCC)

A total of 67 awards were issued by BD2K in these five FOA clusters. In addition, each of the BD2K Centers of Excellence has training components. Because the Centers' training is diverse and mainly focused on training about specific tools developed by the Centers, they are not included in this report.

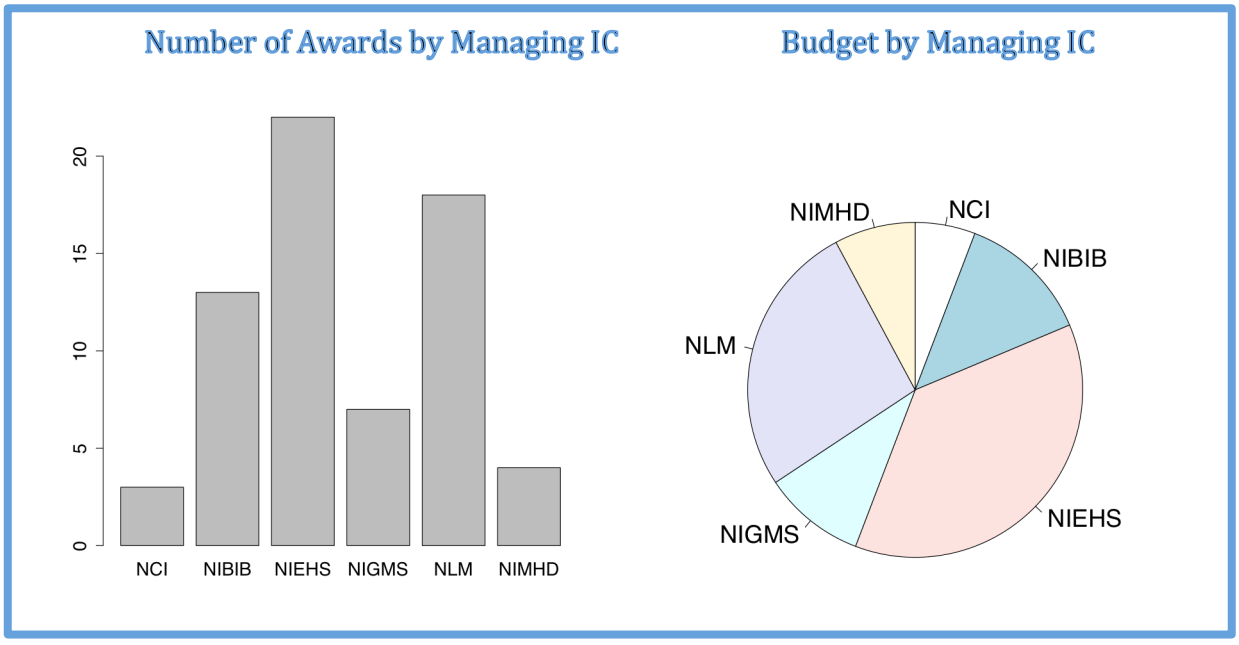
Collectively, the five BD2K training FOA clusters reach an audience of varying experience levels and intentions, from undergraduates to senior faculty and instructors who will be teaching data science. Some of the awards are targeted at particular career levels. For example, the dR25 awards are for undergraduates, the T32/T15 programs are for predoctoral trainees, and the K01 awards are for postdocs and beyond. Other awards, such as the U24 TCC and the R25s are for a broader range of experience levels.



Each cluster addresses one or more goals.



Because BD2K is a trans-NIH program, with funds coming from all ICs and the Common Fund, IC input has been actively solicited. The FOAs were developed by a trans-NIH group of program directors, first called the "BD2K Training Subcommittee" and later referred to as the "BD2K Training Program Management Group". This group has welcomed all-comers, from all ICs and the NIH/OD, with effort to recruit members through two presentations to the TAC (Training Advisory Committee). All programmatic aspects, including FOA development, review attendance, pay plan development, and decisions about award management, have been done through the trans-NIH group that includes 13 ICs, the Common Fund, and the Office of Behavioral and Social Science Research. Day-to-day management of the awards is distributed across 6 ICs in an effort to balance including as many ICs as possible and ensuring that grantees are treated uniformly.



Goal 1: To Improve Big Data Skills of Biomedical Scientists

To improve the big data skills of all biomedical scientists, training opportunities need to be available, and biomedical scientists need to find and access the ones that best fit their needs. To this end, BD2K supports the development of training opportunities and their dissemination to large numbers of learners, as well as infrastructure for discovering them.

To help biomedical scientists find, access, and choose training opportunities, the Training Coordination Center is creating an Educational Resource Discovery Index (ERuDIte). ERuDIte is planned to be a discovery index that organizes pointers to educational content, utilizing metadata describing the educational resource. Utilizing and extending common metadata is being pursued through an international collaboration between the TCC and ELIXIR, a European-based federation of organizations that build infrastructure for the life sciences.

ERuDIte, when combined with a knowledge map that shows how Big Data skills relate to one another, may form the basis of a personalized learning system for biomedical scientists to efficiently acquire new skills to tackle Big Data.

Creation of the content that ERuDIte organizes is supported by BD2K primarily through R25s for open educational resources (OER) and short courses using FOAs HG14-008, HG14-009, LM15-001, and LM15-002.

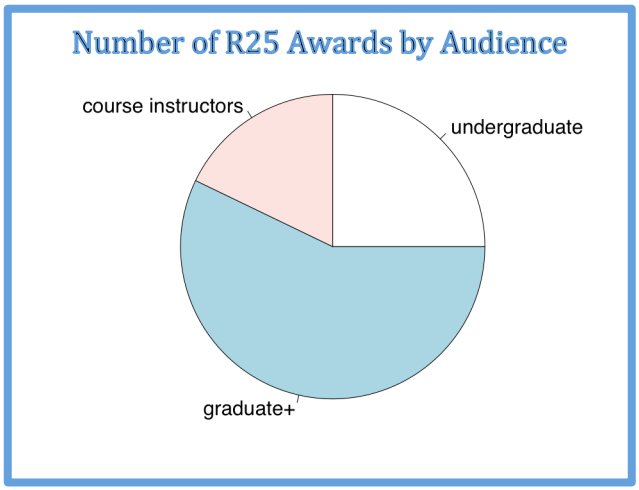
Educational content may also come from other sources, including non-BD2K ones. Some of the BD2K Centers, which each have a training component, are producing educational content such as TED-like talks. Curriculum for data science courses may come from BD2K T32/T15 training programs, which were given the opportunity to apply for $20K in funds to develop and share curriculum of new courses. The number of educational resources supported by BD2K, or even NIH, is dwarfed by the number supported elsewhere, whether by the National Science Foundation, the Department of Education, foundations, universities, or private industry. Educational resources from all of these sources will be included as content in ERuDIte.

The BD2K Open Educational Resources, Short courses, and Diversity programs all aim to introduce biomedical scientists to data management and data science. These programs were designed to be flexible, to allow for innovation and for specialization to a particular audience, domain science, or data science. Although a few of the funded programs are confined to a narrow **audience** or **scientific** area, most are for general audiences and multiple data types. In FY14 and FY15, a total of 29 R25 programs were funded, and these programs have a broad **reach** geographically. More detail about the audience, scientific breadth, and reach follow.

Audience

The R25 awards address a variety of educational levels (undergrads to senior faculty) and intended usages (end users or instructors).

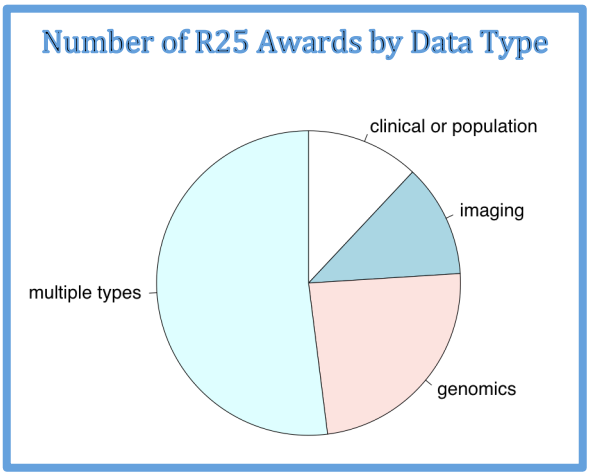
- Five programs focus on helping instructors of advanced undergrad/early grad courses. Examples include:
  - A train-the-trainers course in biomedical data science for instructors, who will collectively develop a curriculum for undergraduates at non-research-intensive colleges
  - Curricular materials (slides, assessments, reading suggestions) that can be used and adapted by other instructors
  - A Toolkit to help librarians teach data management to biomedical scientists
- Seven programs target undergrads directly, through summer programs (including didactic and research experiences) that aim to recruit data science students into biomedical science or to expose biomedical students to data management and data science. Four of the seven programs have a primary goal of enhancing diversity through partnering with BD2K Centers.
- The remaining 16 programs focus directly on the graduate student or the more advanced learner; although the data management and data science material is introductory, because it is new, learners are just as likely to be senior faculty as graduate students.



The R25 programs are a main component of BD2K's diversity efforts. Four undergraduate programs aim to enhance diversity in the biomedical workforce through partnerships between the BD2K Centers of Excellence and low-resourced institutions. The partnerships support the development of curriculum and research experiences for undergraduates and faculty from low-resourced institutions. Collectively, these four programs reach 134 students over the course of 5 years. However, the number of students touched by the improved curriculum and the strengthened faculty is far greater. In addition to the four programs funded explicitly for diversity, other R25s make serious efforts to recruit and train underrepresented minorities. For example, in the first offering, the short course from Oregon Health Sciences University trained 9 URMs out of the 17 total participants. Although the R25 programs form the core of BD2K's diversity efforts, URMs can be supported by BD2K through the T32/T15 training programs, which must have "diversity recruitment and retention plans," and through diversity supplements (PA-15-322).

Scientific Breadth

- Domain Focus: Although the majority of the programs aim for a general biomedical audience, some of them focus on a particular data type. About a quarter of the programs focus on genomics, and another quarter focus on one of imaging, clinical, or population data. The majority use multiple data types and are for a general biomedical audience.



Areas with few applications and hence few funded applications are the clinical and population sciences. Within these areas, this is notable dearth of activity in mHealth. To encourage applications in the clinical and population sciences, BD2K will work closely with relevant Institutes, Centers and Offices, such as NCATS and OBSSR, to ensure that any new funding opportunity announcements contain language to encourage applications in this area and are widely advertised to the appropriate communities.

- Data Management and Data Science: Collectively, the awards span a broad range of topics, including data management, data exploration, data representation, computing, data modeling, and data visualization. Each of these topics can be further broken down in the following way.
    - Data management: reuse of data, data standards, locating and accessing data and tools, organizing and curating data through ontologies and use of metadata
    - Computing: distributed or parallel computing, workflows, programming, algorithms, optimization, and natural language processing
    - Data representation: data structures and databases
    - Data exploration: data munging and preparation, exploratory data analysis
    - Data Modeling: probability, stochastic modeling, introductory statistics, advanced statistics (e.g. multiple testing, dimension reduction), machine learning, experimental design, Bayesian methods, reproducible research, network models
    - Data visualization and communication

  Although collectively the R25 awards span the range above, they are often covered with little depth, and some of the topics are only covered by one Open Educational Resource, and these tend to be the more technical (e.g. optimization and stochastic modeling) or specialized (e.g. ELSI and team science) topics.

Reach

- Eleven programs with in-person components are spread evenly between East coast, West coast, and the middle of the country. Based on the planned enrollments, the 11 programs will serve over 250 participants in the summer of 2016.
    - The three programs that were funded in FY14 held courses in the summer of 2015. They collectively reached undergraduates, PhD students, and faculty from over 30 different universities.
    - Demand for the 2015 in-person courses was high, with reported acceptance rates for the programs with limited slots as being 35% and 14%. Another program gave support to a limited number of students but opened a large auditorium for the course.
- Fourteen of the awards are online Open Educational Resources. These reach a large number of students and instructors, providing a great value per student.
    - For example, about 5,000 students *completed* the first 8 courses in Rafa Irizarry's series of biomedical Data Science MOOCs in the first offering, amounting to about $40 per student based on an NIH investment of $200K (year 1 direct costs).

Although BD2K is supporting the development and discovery of training resources, continued support in the area is needed for a number of reasons: gaps in content coverage exist (e.g. methods for mHealth data, algorithms and optimization methods, advanced statistics, network models); demand for in-person courses continues to exceed supply; different ways of explaining material resonate with different learners; and materials need to be updated to take into account new science and new developments in the understanding of learning.

Goal 2: To Increase the Number of Biomedical Data Scientists

To increase the number of biomedical data scientists, trainees need to gain the appropriate skills, want to work in biomedical science, and have an appropriate place to work. Because all of these trainees have self-selected toward biomedical science, BD2K's focus is on helping trainees get the appropriate skills initially and use those skills in the long run. Trainees may be 1) predoctoral students, who gain foundations in biomedical data science through **T32/T15 training programs**, 2) postdocs/faculty, who are trained in either biomedical science or data science and recognize the need to complement their existing knowledge and skills through **K01 career development awards**, or 3) students or faculty who need specialized training that is unavailable locally but attainable through a **Research Rotation**.

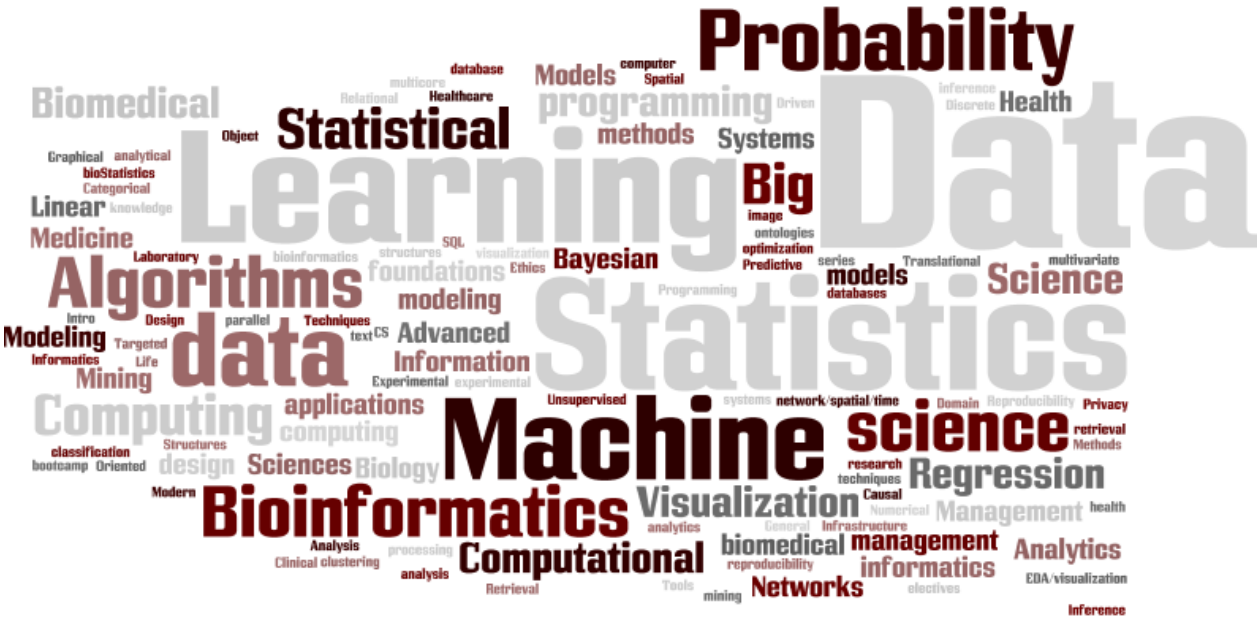
Retaining trainees is both important and a challenge, due to the demand for data science skills across sectors. Although retention is being addressed primarily

through providing support, BD2K is also interested in facilitating conversations with university leaders to discuss career paths for open science and data science. Support for the further development of data scientists into biomedical data scientists is given through K01 awards. Before blending biomedical and data science knowledge, data scientists may be able to immediately contribute to addressing biomedical problems through working in teams with biomedical scientists. Such teams are being fostered through joint NSF/BD2K support in the **QuBBD program**, for building collaborations between data scientists and biomedical scientists.

T32/T15 Training Programs

Predoctoral trainees are supported through new T32 programs and supplements to existing T32 or NLM T15 programs: 6 were funded in FY14, and another 10 will be funded in FY15, for a total of 84 trainees. Since many of the departments that the programs reside in are new, not all programs requested the maximum of 6 slots.

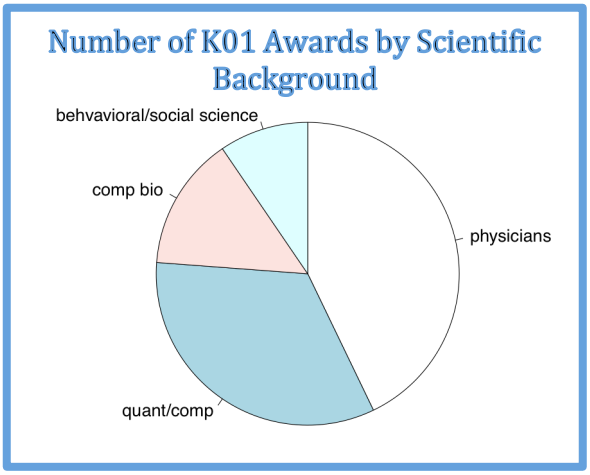
Because the field is nascent, there is still not full agreement as to what the core competencies are. However, some of the common areas of most BD2K training programs are: modern statistics (e.g. handling multiple comparisons, high dimensional data analysis, spatial-temporal correlation), computational techniques (e.g. using cloud and parallel computing, optimization, and algorithms), and machine learning. Most programs are creating new courses in order to integrate these topics together and focus class time on the ones that need to be learned didactically. Below is a word cloud of the core courses of the programs.



K01 Career Development

BD2K is supporting 21 postdocs and faculty with mentored career development awards (K01). The PIs come from diverse backgrounds:

- 9 are physicians, with specialties in hematology/oncology, neurology, neuroradiology, surgery, urologic surgery, pulmonary and critical care medicine, and internal medicine
- 7 have primarily quantitative or computational backgrounds, with degrees in Electrical Engineering and Computer Science, Physics, Nuclear Physics, and Biomedical Engineering
- 3 have backgrounds in fields that blend the biomedical and computational sciences (molecular genetics, bioinformatics and computational biochemistry)
- 2 are behavioral or social scientists (Social Epidemiology, Quantitative Psychology)



The group of K01 awardees is diverse not just by scientific background but also demographically and geographically:

- Nine out of 21 are female.
- They work at 18 unique institutions.

Through a sustained period of research career development and training, the K01 awardees will gain the knowledge and skills necessary to launch independent research careers. The goal is that they become competitive for new research project grant (e.g., R01) funding in the area of Big Data Science.

Research Rotations

The BD2K Training Coordination Center has three main goals: 1) to coordinate by increasing communication across the BD2K-funded training awards; 2) to develop a discovery index for educational resources; and 3) to coordinate research rotations to facilitate access to appropriate expertise. The "research rotations" aim to match trainees who need specialized training with those experts willing and able to provide it. The trainees are expected to mainly be graduate students, postdocs, or junior faculty. Implementation details of the research rotations are still in the development stage, and evaluation measures are being designed along with the program.

NSF/NIH Quantitative Biomedical Big Data (QuBBD) Program

Some problems will require a new model of leadership. Particularly when very diverse skills need to be brought to bear on the problem, teams of individuals with complementary expertise will be needed. Such teams are being fostered through a partnership between NSF and NIH called the QuBBD program. This program consists of a series of Innovation Labs and the funding of planning grants.

Innovation Labs are week-long mentored workshops that catalyze interdisciplinary teams and speed up the process of developing the team's research program. BD2K ran a pilot Innovation Lab in July 2015. By the end of the week, 12 new interdisciplinary teams were prepared to submit grant applications together. The teams submitted applications for small planning grants, along with newly-formed teams that did not go through the Innovation lab.

Because an established team can do much work virtually, the teams fostered by the QuBBD program draw in a wide range of talent, many of whom are from schools not otherwise represented within BD2K. These teams include some data scientists who are in physically isolated locations along with biomedical scientists who are unable to find data science collaborators due to the high demand.

The Summer 2015 Innovation Lab was successful, and the key to success was the participation of mentors who are from a variety of backgrounds and are leaders in their respective fields. The mentors guide teams through the iterative ideation process, offering feedback on research ideas throughout the week. Continuation of the QuBBD program is under discussion.

Summary

The BD2K programs in training described in this document are early efforts to address a need identified by the Advisory Committee to the Director's Data and Informatics Working Group. They cover both biomedical data science specialists, as well as specialists in other biomedical areas. They also cover the educational pipeline from undergraduates to faculty.

These programs are "early efforts" because there is still much work to be done to meet both goals. To quickly increase the base data science skills of a wide variety of biomedical scientists, early resources focused on broad, generally applicable topics. Later resources might be on more specialized topics. Likewise, to quickly jump start the increase in the number of biomedical data scientists, some of the training programs are modifications of existing programs, building on existing infrastructure and courses. As the community converges on the core competencies of the field, biomedical data science training programs will likely evolve and may end up bearing little resemblance to the programs initially funded.

The BD2K Initiative is in its infancy, and the training programs, along with other BD2K programs, are contributing to the development of the field of biomedical data science in the US and across the world. Although the awarded grants are confined to US institutions, the reach extends far beyond through the development of open educational resources and contributions to the global conversations about the field of biomedical data science. In addition, international collaborations surrounding the discovery of open educational resources for biomedical data science have begun. The BD2K program aims to improve the ability of the biomedical workforce to use Big Data both today and tomorrow, both in the US and across the world.

Geographic Distribution of BD2K Training Awards